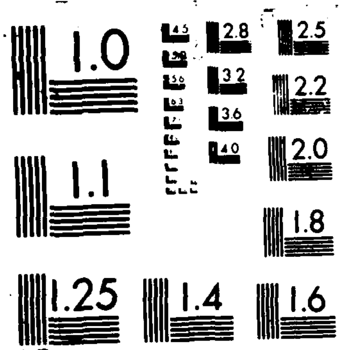
AD-A187 063 MATHEMATICAL STRUCTURES FOR SPEECH RECOGNITION(U) ROME AIR DEVELOPMENT CENTER GRIFFISS AFB NY W TODD JUL 87 RADC-TR-87-109 1/1

UNCLASSIFIED F/G 25/4 NL


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

2

RADC-TR-87-109
In-House Report
July 1987

MATHEMATICAL STRUCTURES FOR SPEECH RECOGNITION

Wayne Todd, Capt, USAF

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, NY 13441-5700

This report has been reviewed by the RADC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-87-109 has been reviewed and is approved for publication.

APPROVED: Thadeus J Domurat

THADEUS J. DOMURAT
Chief, Signal Intelligence Division
Directorate of Intelligence & Reconnaissance

APPROVED: Carlo P Crocetti

CARLO P. CROCETTI
Technical Director
Directorate of Intelligence & Reconnaissance

FOR THE COMMANDER: Richard W. Pouliot

RICHARD W. POULIOT
Directorate of Plans & Programs

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

ADA187063

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b RESTRICTIVE MARKINGS N/A
2a SECURITY CLASSIFICATION AUTHORITY N/A	3 DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b DECLASSIFICATION/DOWNGRADING SCHEDULE N/A	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) RADC-TR-87-109	5 MONITORING ORGANIZATION REPORT NUMBER(S) N/A
6a NAME OF PERFORMING ORGANIZATION Rome Air Development Center 6b OFFICE SYMBOL (If applicable) IRAA	7a NAME OF MONITORING ORGANIZATION N/A
6c ADDRESS (City, State, and ZIP Code) Griffiss AFB NY 13441-5700	7b ADDRESS (City, State, and ZIP Code) N/A
8a NAME OF FUNDING/SPONSORING ORGANIZATION Rome Air Development Center 8b OFFICE SYMBOL (If applicable) IRAA	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N/A
8c ADDRESS (City, State, and ZIP Code) Griffiss AFB NY 13441-5700	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 61102F PROJECT NO: 2305 TASK NO: J8 WORK UNIT ACCESSION NO: 32

11 TITLE (Include Security Classification)
MATHEMATICAL STRUCTURES FOR SPEECH RECOGNITION

12 PERSONAL AUTHOR(S)
Wayne Todd, Capt, USAF

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
In-House	FROM Apr 85 TO Dec 86	July 1987	32

16 SUPPLEMENTARY NOTATION
N/A

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP
20	01	

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Speech Recognition
Mathematical Intelligence (MI)

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

The Mathematical Intelligence (MI) Development System was implemented on a DEC PDP 11/70 computer, an FPS 5210 array processor, and a Ramtek 9465 color graphics display. The two major components of the system are the Digital Speech Processing Tool and the MI Speech Recognizer. The MI Speech Recognizer was informally tested for pitch extraction and results were encouraging. Additional work is necessary to formally test the MI Speech Recognizer and complete the MI Development System.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a NAME OF RESPONSIBLE INDIVIDUAL Wayne Todd, Capt, USAF	22b TELEPHONE (Include Area Code) (315) 330-4024 22c OFFICE SYMBOL RADC (IRAA)

DD Form 1473, JUN 86 Previous editions are obsolete. SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

PREFACE

Part of the Rome Air Development Center's (RADC) speech research is an in-house effort to develop vector classification concepts for recognition problems in general and for speech recognition in particular. These concepts, referred to as Mathematical Intelligence (MI), were developed by Capt William I. Lundgren prior to his departure from RADC in October 1985. This report is designed to familiarize the reader with MI, show how it applies to speech recognition, and discuss the preliminary test results.

The MI system developed at RADC consists of numerous programs that were written to run on a DEC PDP 11/70 minicomputer and an FPS 5210 floating point array processor. This configuration is discussed in more detail in Chapter 3. Many of the programs implement a rudimentary digital speech processing tool (DSPT) that performs such functions as digital record and playback of speech, speech spectrogram display, and symbolic labeling of spectrogram segments. This tool is useful for observing speech data graphically to build the MI structure manually as discussed in Chapter 2. The remainder of the programs implement MI speech recognition. At this time, the DSPT is functional and available for use. The MI speech recognizer is not.

The numerous MI programs developed at RADC are, for the most part, undocumented. In addition, the algorithms used to write the programs are also undocumented. For example, there is no documentation on the windowing algorithm used in the DSPT. The author believes, however, that the MI concepts are sufficiently documented in this report to allow any speech researcher to implement an MI development system (Chapter 2).

TABLE OF CONTENTS

1. Introduction

The goal of speech recognition is to provide a more natural way for humans to enter commands and data to a computer. In some cases, this is a convenience. In other cases, where a human's hands and/or eyes are already busy, speech input is a necessity. The success of applying current speech recognition technology is largely dependent upon the particular application because human-like performance has not yet been achieved. Some of the key issues in speech recognition performance are listed in Table 1-1. Also shown is how these issues compare between a fragile system and a robust system. A fragile system's performance deteriorates quickly when strict constraints are not maintained. A robust system requires fewer constraints. Humans are the most robust speech recognition system available today.

Two of the most popular techniques currently used for speech recognition are Dynamic Time Warping (DTW) and Hidden Markov Modeling (HMM). A detailed discussion of these techniques is beyond the scope of this report. However, the reader who is not familiar with these techniques can consult the references. Since these techniques have not produced human-like performance, further research and development are required to both modify these techniques and to investigate unique alternatives. MI falls into the later category.

Table 1-1 Speech Recognition Issues

Issue	Fragile System	Robust System
Continuous/ Isolated Speech	Isolated and Limited Connected	Continuous
Intraspeaker Variability - Stress - Inflections - Enunciation - Vocal Tract Changes (e.g., Illness)	Increase in Recognition Errors	Accurate Recognition
Interspeaker Variability - Male/Female - Dialect	Increase in Recognition Errors	Accurate Recognition
Environmental Variability - Transducer Response - Noise - Acoustic Chamber	Increase in Recognition Errors	Accurate Recognition
Vocabulary Size	Small	Large
Speech Understanding	None or Syntactic Only	Syntactic, Semantic, Pragmatic, & Goal Understanding
Computational Power Required	Small	Large

2. Mathematical Intelligence

Both methods of recognition mentioned in the previous chapter attempt to map similar inputs into a single class so a particular class can be recognized. Similarity is measured between expected (training) data and actual (recognition) data. Various methods are used to modify the input so that it will be similar enough to match at least one class. If the inputs are not similar enough, recognition will not occur. With this approach, all of the input information is used to determine similarity of the inputs and subsequent vector classification. The key distinguishing feature of Mathematical Intelligence (MI) is that information in the input not relevant to the recognition is filtered out by means of a mathematical structure. In other words, not all of the input information is used to classify a particular vector.

2.1. Vector Classification Concepts

This section describes how MI maps an input feature vector into its appropriate class (Figure 2.1). The input feature vector is derived from pre-determined mathematical operations on an input analog signal. The mathematical operations result in parameters that describe various features of the analog signal. These feature vectors are sequenced in time so that each feature vector describes the analog signal for some specific period of time. That period of

time is called a frame or window. A feature vector can contain features that span more than one window.

Once a feature vector is determined, the magnitudes of the features within that vector are compared to each other. This is accomplished using pre-defined pseudo-logical functions. For example, let the values of features A, B, C, and D be 0.8, 0.95, 0.2, and 0.54, respectively. Then a function, F1, can be described as "A and B are large compared to C." MI obtains this example relationship mathematically by

$$F1 = \frac{A \times B}{1 + C}$$

The number and type of functions used is determined by what is required to separate an input feature vector into its class and by the limit of information contained in that vector. The combined set of pre-defined functions forms the function vector which is applied to each class.

For each class, there is a Threshold Vector where each component corresponds to a component in the Function Vector. If the function component is greater than the corresponding threshold component, then a value of one is assigned; otherwise a value of zero is assigned. The result of this sequence of operations is a pre-output vector for each class composed of a set of ones and zeros.

Next, the ones and zeros in each Pre-output Vector are summed to obtain an integer value corresponding to each class. The class with the largest integer value is the selected class for the input feature vector.

The MI vector classification procedure can be described mathematically as follows:

Pm = TH(F1,T1m), TH(F2,T2m), ... , TH(Fn,Tnm)

Class = NUM (MAX (I1,I2,...,Im))

where

P = pre-output vector (one for each class).

m = class number.

TH = threshold function; results in a one if the first argument is greater than the second, otherwise results in a zero.

F = function.

n = function number.

T = threshold.

Class = class that the input feature vector is mapped into.

MAX = function that determines the maximum value of its arguments.

I = integer sum of ones and zeros of the pre-output vector.

NUM = function that obtains the class number associated with a particular integer sum.

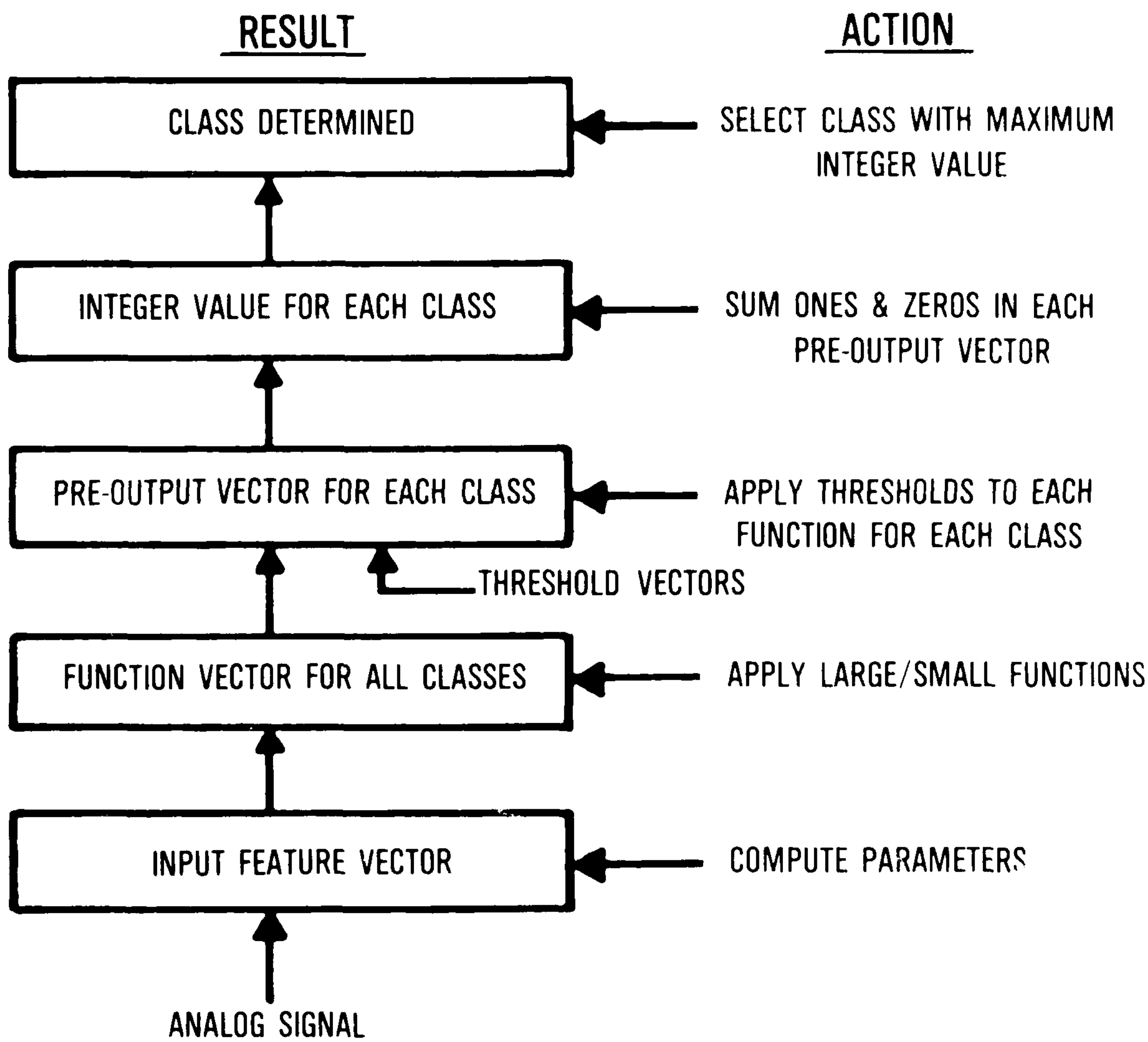


Figure 2.1 Mapping of an Input Feature Vector into its Class

2.2. Vector Classification Example

Suppose there is an Input Feature Vector with four features, and the objective is to map it into one of six classes. The following are given:

Input Feature Vector = (A, B, C, D)

Feature Values: A = 0.25, B = 0.87, C = 0.79, D = 0.62

Function Vector = (F1, F2, F3, F4, F5)

Functions:

$$F1 = \frac{A \times B \times C}{1 + D}$$

$$F2 = \frac{A \times D}{(1 + B)\,(1 + C)}$$

$$F3 = \frac{B \times C}{(1 + A)\,(1 + D)}$$

$$F4 = \frac{C}{(1 + A)\,(1 + B)\,(1 + D)}$$

$$F5 = \frac{B}{(1 + A)\,(1 + C)\,(1 + D)}$$

Threshold Vectors:

Class 1: (.33, 1.0, .41, 1.0, 1.0)

Class 2: (1.0, .25, 1.0, .39, 1.0)

Class 3: (1.0, 1.0, .27, 1.0, .43)

Class 4: (.44, 1.0, 1.0, .36, 1.0)

Class 5: (1.0, .35, 1.0, 1.0, .26)

Class 6: (1.0, 1.0, .35, .42, 1.0)

By applying the functions of the Function Vector to the Input Feature Vector, the Evaluated Function Vector (EFV) obtained is

EFV = (.11, .05, .34, .21, .24).

Next, a Pre-Output Vector is calculated for each class. For Class 1,

P1 = (TH(F1,T11), TH(F2,T21), TH(F3,T31), TH(F4,T41), TH(F5,T51))

= (TH(.11,.33), TH(.05,1.0), TH(.34,.41), TH(.21,1.0), TH(.24,1.0))

= (0, 0, 0, 0, 0)

Thus,

I1 = 0.

Similarly, for Class 2 - 6,

P2 = (0, 0, 0, 0, 0) I2 = 0

P3 = (0, 0, 1, 0, 0) I3 = 1

P4 = (0, 0, 0, 0, 0) I4 = 0

P5 = (0, 0, 0, 0, 0) I5 = 0

P6 = (0, 0, 0, 0, 0) I6 = 0

Then,

Class = NUM (MAX (I1, I2, I3, I4, I5, I6))

= NUM (MAX (0, 0, 1, 0, 0, 0))

= NUM (1)

= 3

Thus, the Input Feature Vector, (.25, .87, .79, .62) belongs to Class 3. This procedure is repeated for every Input Feature Vector.

2.3. Speech Recognition

The MI vector classification procedure is directly applicable to speech recognition. In this case, the analog signal (Figure 2.1) is transduced speech. The speech is digitized and various parameters are extracted to construct the Input Feature Vector (IFV). Features which might be considered include the Power Spectral Density for a given set of frequency bands, pitch period, formant frequencies, etc. The reader can consult the references for a more detailed discussion of speech features.

Recognition occurs when the IFV(s) is mapped into its appropriate class. Thus, there is one class or sequence of classes that corresponds to each speech segment that the system has been trained to recognize.

2.4 Training

Training an MI speech recognizer to recognize speech segments involves constructing the Function Vector for all classes and a Threshold Vector for each class. In the example of MI vector

classification presented in an earlier section, the Function Vector and the Threshold Vectors were simply given. However, these vectors must somehow be derived by training the system. This can be done manually by the user or automatically by the system.

It is crucial in the training process that valid selection criteria are used to select the Function Vector and Threshold Vectors that will be used for recognition. Unfortunately, the author is not familiar with the criteria that were used in implementing the MI system at RADC, nor the results of testing those criteria.

2.4.1. Manual Structure Building

In order for the user to manually build an MI structure, he must be able to isolate the speech segment to be recognized and observe the sequence of Feature Vectors that represent that segment. The user can then determine the relationships between the normalized values (from 0 to 1) of the features, based on some selection criteria.

For example, suppose the problem is to recognize several spoken words, and the training vocabulary consists of the same words. An algorithm could be employed to calculate the power spectral density (PSD) values for a spectrum divided logarithmically into 32 frequency bands. Each Feature Vector then would have 32 features.

These features can be displayed as a spectrogram on a high-resolution color graphics device. The color represents the PSD of each feature over a single frame. The user observes the colors to compare relative PSD between features and derives functions and Threshold Vectors that describe the relationships. The choice of functions and Threshold Vectors must be such that each word in the training vocabulary is unique.

2.4.2. Automatic Generation

In the example presented in the previous section, the user was required to select functions and Threshold Vectors. For a vocabulary of just a few words (or other speech segments) the selection criteria need not be well-defined as long as acceptable recognition performance is achieved. However, for a vocabulary of practical size (100+ words), the selection criteria must be highly definitive, or the user will spend endless hours "tweaking" the system to achieve acceptable recognition performance. If the selection criteria can be well-defined, then it's also possible to automate the selection of functions and Threshold Vectors. This is the essence of the automatic generation process.

2.5. MI Development System

In order to develop an MI speech recognizer within reasonable

time, cost and manpower constraints, a complete development system is required. The essential components and their interrelationships are illustrated in Figure 2.2. The central component is the Digital Speech Processing Tool (DSPT).

The DSPT interacts directly with the user interface and with all other components of the MI development system. It has facilities for evaluating recognition results, developing DSP algorithms, training the system either manually or automatically, and supplying both raw and derived data to the user interface. DSP algorithms, the Function Vector, and the Threshold Vectors originate in the DSPT and are then down-loaded into the MI recognizer. DSPT capabilities are summarized in Table 2-1.

TABLE 2-1 DSPT Capabilities

Digital Record/Playback	DSP Feature Extraction
Provide Raw Data to User Interface	Provide Feature Data to User Interface
Window Type	Window Length
Window Overlap	Spectral Shaping
High Pass Filter	Low Pass Filter
Band Pass Filter	Sampling Rate
Sample Quantization	Length of Speech Segment
Various Display Options	DSP Algorithm Development

The Speech Data Base consists of digitally recorded audio correlated with the symbolic representation (e.g., phonemes, words) of the audio. This is referred to as a "labeled" speech data base. Having correlated audio data and symbolic data allows for a direct automated evaluation of the recognition results. It also allows the automatic generation algorithm to train the system without user intervention.

The User Interface consists of various Input/Output devices that link the user to the DSPT and the MI Speech Recognizer. These include microphones, speakers, amplifiers, textual display, graphics display, keyboard, etc.

2.6. Unresolved Difficulties

The theoretical basis for MI speech recognition can be found in two of the references (Lundgren 84, Lundgren 85). Although it is not based on any well-recognized model of human aural cognition, it can still be useful. The author has identified several possible problems that should receive additional analysis and that may need to be resolved to make the MI method more useful.

2.6.1. Selecting Largest Integer

Part of the example of MI vector classification presented in an earlier section was to determine the integer sums (one for each class), determine the maximum of the integer sums, and select the class corresponding to the maximum integer sum. A conflict can arise if the maximum integer sum value belongs to more than one class. Of course,

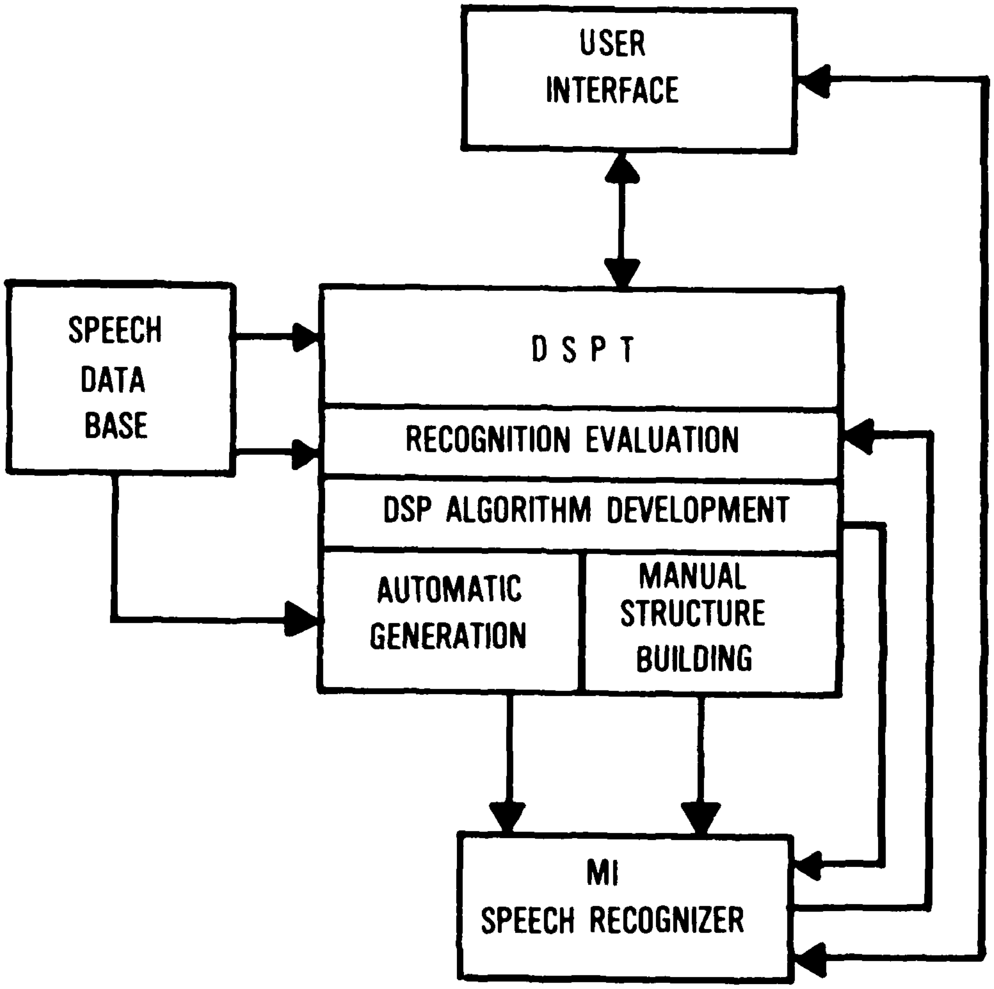


Figure 2.2 MI Speech Recognition Development System

the Function Vector and Threshold Vectors constructed during training should minimize the possibility of such a conflict. Nevertheless, it is possible and should be dealt with. The author is not a ire of whether or not this issue was addressed in the RADC MI system.

2.6.2. Non-Linear Functions

When a particular feature of an Input Feature Vector is used in the numerator of a function, the calculated function value will change linearly with respect to changes in the feature value. However, if the feature is used in the denominator of a function, the calculated function will change non-linearly with respect to changes in the feature value. The author does not know why this dichotomy of functional behavior was implemented or whether it poses a problem in the recognition process.

Consider as an example a function, F1=(AxB)/C and a sequence of three Input Feature Vectors,

(A = 0.85, B = 0.75, C = 0.20)

(A = 0.85, B = 0.75, C = 0.40)

(A = 0.85, B = 0.75, C = 0.80).

The calculated values for F1 are 0.53, 0.46 and 0.35, respectively. While the value of C increased by a factor of two in each case (vector 1 to 2, vector 2 to 3), the value of F1 did not decrease by a factor of two. In fact, F1 first decreased by a factor of 0.87 and next by a factor of 0.76, indicating a non-linear relationship between the value of C and F1.

2.6.3. Combining Information

Suppose the values of an Input Feature Vector are (A = 0.75, B = 0.87, C = 0.98, D = 0.77). Two valid functions that could be constructed and correspond to the feature values are F1=AxBxCxD and F2=AxB. Intuitively, it seems because both functions apply, and because F1 uses more information from the Input Feature Vector than F2, that the calculated value of F1 would exceed F2. In other words, one should have more confidence in F1 and F2. Yet just the opposite is true. The confidence that A, B, C, and D are large is 0.49 for F1 and 0.65 for F2, even though F2 considered only two features! This reverse logic says, in effect, "to gain the highest confidence that a particular condition exists, use the least amount of information."

3. Initial Testing of MI

Although speech recognition using MI has not been formally tested, some informal testing has been done. More specifically, pitch detection was tested using MI. The problem is to correctly classify the pitch frequency of a speaker's voice within each 10 Hz class for the duration of the speech utterance examined. The automatic generation algorithm was not used so that the MI structure itself and not the automatic generation could be evaluated.

3.1 System Configuration

Figure 3.1 is a block diagram of the basic system used to develop the MI software and perform the initial testing. The operating system was RSX-11M by Digital Equipment Corporation (DEC). The software run on the PDP-11 was written in Fortran 77. Software was also written to run on the Floating Point Systems (FPS) 5210 array processor.

3.2 Pitch Detection

The input data to the pitch recognition system was generated by sampling speech at a rate of 10 kHz and using a Fast Fourier Transform (FFT) on 1024 samples to generate a 512-point Power Spectral Density (PSD). This corresponds to examining speech segments 102.4 msec in length. A Hanning window was used with a weighting of 6 dB/octave above 500 Hz. The pitch information is

contained in pitch peaks that are located in the fundamental (the pitch itself) and the harmonics (integral multiples of the pitch). Informal test results indicated that the MI system identified the correct pitch about 90% of the time.

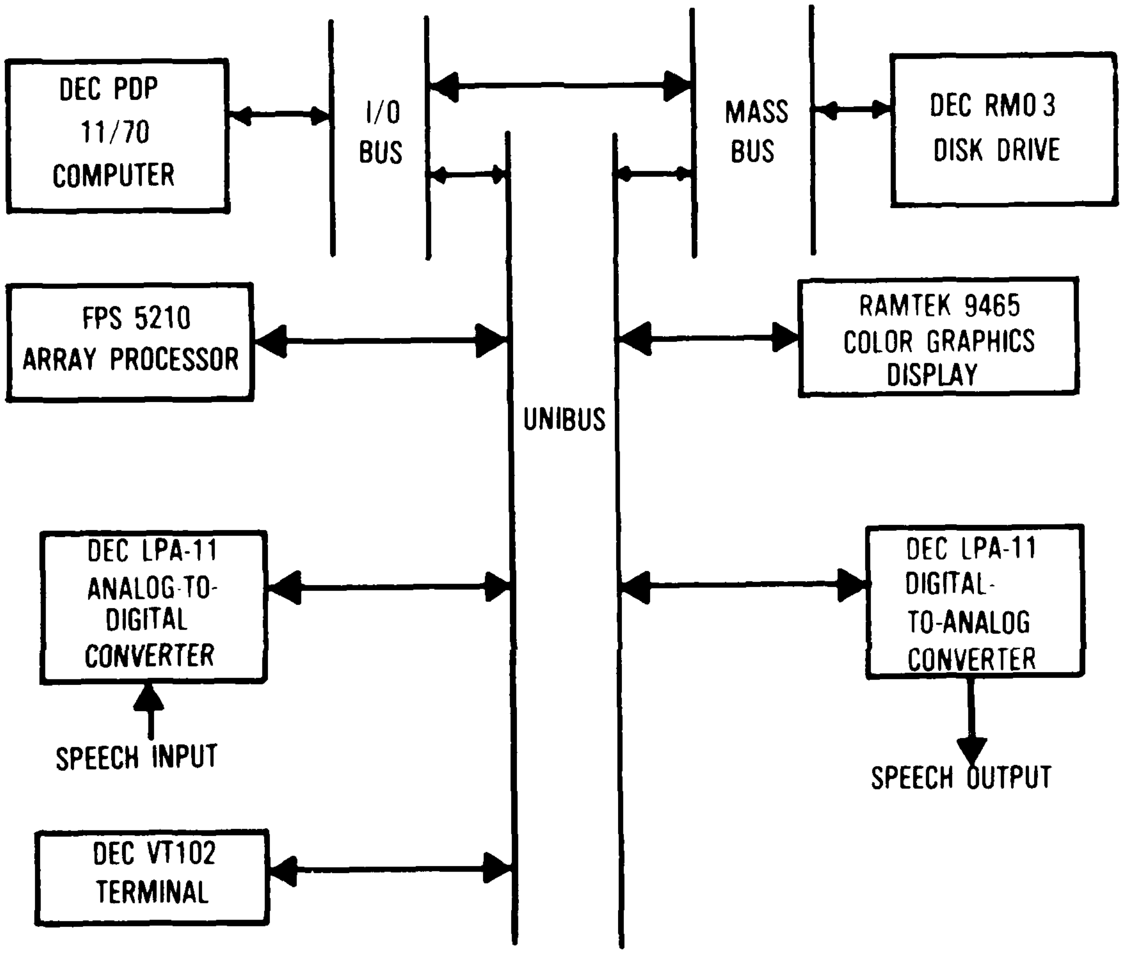


Figure 3.1 Block diagram of the MI development system hardware

4. Summary and Conclusions

This report has discussed the problem of speech recognition and how MI can be employed to solve the problem. The basic MI concepts were described in enough detail to allow a speech researcher to implement one form of a MI system. Much research is necessary to permit a meaningful comparison of MI to other speech recognition techniques. This will require a well-integrated MI development system as described earlier and a library of popular speech data bases used to evaluate speech recognition. Currently, only a partial development system exists at RADC and there are no plans to extend it beyond its present form.

5. Bibliography

Bahl, L. R. et al "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Transactions on Pattern Recognition and Machine Intelligence, Vol. PAMI-5, pp. 179-190, 1983.

Itakura, F. "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-23, pp. 67-72, Feb. 1975.

Juang, B. H. "On Hidden Markov Model and Dynamic Time Warping for Speech Recognition - A Unified View," AT&T Bell Lab. Tech Journal, Vol. 63, No. 7, pp. 1213-1243, Sep. 1984.

Levinson, S. E. and M. Y. Liberman "Speech Recognition by Computer," Scientific American, 244: 64-76, April 1981.

Lundgren, W. I. "Mathematical Intelligence: A Method of Optimizing Vector Classification," Proceedings of Speech Tech 84, pp. 81-86.

Lundgren, W. I. "Mathematical Intelligence: The Automatic Generation Concepts," Proceedings of Speech Tech 85, April 1985.

Mari, J. F. and S. Roucos "Speaker Independent Connected Digit Recognition Using Hidden Markov Models," Proceedings of Speech Tech 85, April 1985.

Pallett, David S. "Performance Assessment of Automatic Speech Recognizers," Journal of Research of the National Bureau of Standards, Vol. 90, No. 5, Sep - Oct 1985.

Rabiner, L. R. and B. H. Juang "An Introduction to Hidden Markov Models," IEEE ASSP Magazine, Vol. 3, No. 1, Jan. 1986.

Reddy, D. R. "Speech Recognition by Machine: A Review," Proceedings of the IEEE, pp. 501-531, April 1976.

Sakoe, H. and S. Chiba "A Dynamic Programming Approach to Continuous Speech Recognition," Proceedings of the 7th International Congress on Acoustics, Budapest, Hungary, Paper 20C-13, Aug. 1971.

Sakoe, H. and S. Chiba "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-26, pp. 43-49, Feb. 1978.

MISSION
of
Rome Air Development Center

RADC plans and executes research, development, test and selected acquisition programs in support of Command, Control, Communications and Intelligence (C^3I) activities. Technical and engineering support within areas of competence is provided to ESD Program Offices (POs) and other ESD elements to perform effective acquisition of C^3I systems. The areas of technical competence include communications, command and control, battle management, information processing, surveillance sensors, intelligence data collection and handling, solid state sciences, electromagnetics, and propagation, and electronic, maintainability, and compatibility.

END
Feb.
1988
DTIC